Bioinformatics, 2024, **40(11)**, btae613
https://doi.org/10.1093/bioinformatics/btae613
Advance Access Publication Date: 16 October 2024
Original Paper

OXFORD

Data and text mining

Lifestyle factors in the biomedical literature: an ontology and comprehensive resources for named entity recognition

Esmaeil Nourani [1,2], Mikaela Koutrouli [1], Yijia Xie [1,4], Danai Vagiaki [1,5], Sampo Pyysalo [3], Katerina Nastou [1,*], Søren Brunak [1,*], Lars Juhl Jensen [1,*]

[1]Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Copenhagen 2200, Denmark
[2]Faculty of Information Technology and Computer Engineering, Azarbaijan Shahid Madani University, Tabriz, Iran
[3]TurkuNLP Group, Department of Computing, Faculty of Technology, University of Turku, Turku 20014, Finland

[4]Present address: Zhejiang Ponshine Information Technology Co., Ltd., Hangzhou 311100, China
[5]Present address: Genome Biology Unit, European Molecular Biology Laboratory (EMBL), Heidelberg, Germany; Collaboration for joint PhD degree between EMBL and Heidelberg University, Faculty of Biosciences, Heidelberg, Germany; Division of Computational Genomics and Systems Genetics, German Cancer Research Center (DKFZ), Heidelberg, Germany

*Corresponding authors. Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Copenhagen 2200, Denmark. E-mail: katerina.nastou@cpr.ku.dk (K.N.); soren.brunak@cpr.ku.dk (S.B.); lars.juhl.jensen@cpr.ku.dk (L.J.J.)

Associate Editor: Zhiyong Lu

Abstract

Motivation: Despite lifestyle factors (LSFs) being increasingly acknowledged in shaping individual health trajectories, particularly in chronic diseases, they have still not been systematically described in the biomedical literature. This is in part because no named entity recognition (NER) system exists, which can comprehensively detect all types of LSFs in text. The task is challenging due to their inherent diversity, lack of a comprehensive LSF classification for dictionary-based NER, and lack of a corpus for deep learning-based NER.

Results: We present a novel lifestyle factor ontology (LSFO), which we used to develop a dictionary-based system for recognition and normalization of LSFs. Additionally, we introduce a manually annotated corpus for LSFs (LSF200) suitable for training and evaluation of NER systems, and use it to train a transformer-based system. Evaluating the performance of both NER systems on the corpus revealed an *F*-score of 64% for the dictionary-based system and 76% for the transformer-based system. Large-scale application of these systems on PubMed abstracts and PMC Open Access articles identified over 300 million mentions of LSF in the biomedical literature.

Availability and implementation: LSFO, the annotated LSF200 corpus, and the detected LSFs in PubMed and PMC-OA articles using both NER systems, are available under open licenses via the following GitHub repository: https://github.com/EsmaeilNourani/LSFO-expansion. This repository contains links to two associated GitHub repositories and a Zenodo project related to the study. LSFO is also available at BioPortal: https://bioportal.bioontology.org/ontologies/LSFO.

1 Introduction

Lifestyle is a complex and multifaceted concept that involves nutrition, behaviors, habits, and activities that individuals engage in, which can impact their health. These lifestyle factors (LSFs) are widely recognized as important in shaping individual health trajectories and influencing the onset and progression of diseases (WHO 2023). Numerous studies illustrate the association between an unhealthy lifestyle and an elevated risk of chronic disease as well as the potential of reducing the risk by adopting a healthy lifestyle (Subramanian *et al.* 2020, Nyberg *et al.* 2020, Tobias *et al.* 2023). LSFs can thus play a pivotal role together with genetic factors in precision medicine (Jeon *et al.* 2018, Gray *et al.* 2020, Gabbert *et al.* 2023, Yurkovich *et al.* 2024), and a crucial first step is to structure the existing information on LSFs.

Specific types of LSFs such as environments, exposure to risk factors, smoking, and food are covered by existing resources like the Environment Ontology (Buttigieg *et al.* 2013), the Exposome Explorer (Neveu *et al.* 2020), the Cigarette Smoke Exposure Ontology (Younesi *et al.* 2014), and FoodOn (Dooley *et al.* 2018). However, these fall short in comprehensively covering other crucial disease-associated LSFs. Since, there is no clear definition in the literature of what constitutes an LSF, we define LSFs to be all non-genetic health determinants associated with diseases. This includes categories such as physical and leisure time activities, socioeconomic factors, personal care products and cosmetic procedures, sleep, mental health practices and substance use, on top of the categories mentioned above. The fact that there is no comprehensive ontology covering all these categories severely hampers formalization of information on LSFs, including the development of text-mining solutions to help extract it from literature.

The first critical step to alleviate this issue is to develop a named entity recognition (NER) method to identify LSF mentions within scientific publications. Many such systems exist for other biomedical entities [reviewed by Huang *et al.*

Received: 22 May 2024; **Revised:** 26 September 2024; **Editorial Decision:** 6 October 2024; **Accepted:** 15 October 2024

(2020), Perera *et al.* (2020), Song *et al.* (2021)]. Developing an NER method for LSFs requires either a comprehensive LSF dictionary that can be matched against text to find them or a text corpus with manually annotated LSFs that can be used to train deep learning-based methods.

In this article, we address the task of recognizing LSFs within biomedical text by presenting four major contributions. Firstly, we introduce a novel LSF ontology (LSFO), featuring a multilevel hierarchical structure. It includes the main LSF categories at the top level and extends to specific subcategories and low-level concepts. Secondly, we introduce the first annotated corpus for LSFs, LSF200, a valuable resource for the BioNLP community as it can be used for training and evaluating NER systems. Thirdly, we introduce a dictionary-based NER system enabling for the first time the detection of a diverse set of lifestyles in the literature. This NER system leverages the LSFO for dictionary creation, thus enabling both recognition and normalization for the matches to LSFO concepts. Lastly, we introduce a transformer-based NER system (Vaswani *et al.* 2017) for LSF detection, leveraging LSF200 for training and evaluation.

2 Materials and methods

Currently, there is a lack of resources to capture the vast diversity of LSFs, under a single umbrella. Below, we have made an effort to generate a comprehensive categorization of the different aspects of LSFs accompanied by a small description of concepts that fit in each category. The creation of this categorization is the first crucial step that allowed us to annotate a text corpus and create a dictionary of LSFs, for the purposes of deep learning-based NER and dictionary-based NER, respectively. In the next sections, we provide details in the methodology used to achieve our goals.

2.1 The LSF categorization

Within the context of lifestyle, we have identified nine categories that can collectively describe all LSFs, namely:

1) *Nutrition*, a category that covers different branches such as dietary habits, food groups, food processing and preparation, macronutrients, and micronutrients, among others.
2) *Socioeconomic factors*, which includes social and economic conditions such as income, wealth, education, and socioeconomic status.
3) *Environmental exposures*, includes exposure to various environmental factors, such as air pollution, water quality, and workplace hazards, that can impact an individual's health.
4) *Substance use*, covers concepts such as smoking, as well as illicit drug use.
5) *Physical activities*, includes regular exercise and physically demanding activities like leisure time, occupational, and household physical activities.
6) *Non-physical leisure time activities*, describes any activity an individual might engage in during their free time that is not a physical activity.
7) *Personal care products and cosmetic procedures*, covers activities related to hygiene, use of cosmetic and cleaning products as well as invasive procedures that people undergo to improve their appearance, such as cosmetic surgery.
8) *Sleep*, covers sleep quality, stages, and habits.
9) *Mental health practices* includes the behaviors and habits related to maintaining good mental health and emotional well-being such as meditation, and psychotherapy.

2.2 Annotation of the LSF200 text corpus

To create an LSF text corpus, we selected three of the most relevant journals within each of the nine LSF categories introduced above. We selected 200 abstracts for our corpus, evenly distributed across all categories. We aimed for a balanced selection of documents among LSF categories, considering the chosen journals' scope and focus. Selected top three journals per category are available in Supplementary Section S1. To ensure high diversity, we opted to annotate abstracts instead of full-text documents, as this would allow us to annotate a larger number of documents and thus obtain a wider selection of different entities with the same curation effort.

The annotation process started by creating an initial set of annotation guidelines, which we improved through two rounds of refinement. For each of the three versions of the guidelines, two annotators annotated a new set of 15 abstracts, based on which we calculated the inter-annotator agreement. A meeting was held after each round to discuss disagreements, update the guidelines, and clarify any ambiguities or gaps in the rules that caused the disagreements between the annotators. We subsequently annotated the entire LSF200 according to the final guidelines (https://esmaeilnourani.github.io/lifestylefactors-annotation-docs/entities), with each annotator working on a different set of abstracts. We used the BRAT rapid annotation tool for document annotation (Stenetorp *et al.* 2012).

2.3 Dictionary construction

One common approach to NER is to develop a dictionary-based system (Cook and Jensen, 2019), leveraging existing databases or ontologies to create the dictionary. For instance, a disease NER system can be established by constructing a dictionary derived from existing resources like the Disease Ontology (Baron *et al.* 2024). While there are existing ontologies that cover some aspects of LSFs, there is no comprehensive resource that encompasses all aspects; thus, we manually constructed an initial dictionary inspired by existing biomedical literature, ontologies, and questionnaires. The aim was to come up with a wide selection of names belonging to each of the nine categories of LSFs described above, which would serve as a good starting point for semi-automatic expansion. The guidelines for the creation of the manual version of the dictionary are provided in Supplementary Section S2.

Then, we wanted to introduce names from a resource that would contain a diverse set of candidates. As 45% of the names in the initial dictionary exist as Wikipedia pages, we decided to use Wikipedia page titles as a source of potentially missing LSFs. To predict which page titles are good LSF candidates, we used a transformer-based approach that scores names based on their contexts in biomedical literature (Nastou *et al.* 2023), and manually assessed all high-scoring candidates before adding them to the dictionary. For more details on the training, evaluation and prediction process please refer to Supplementary Section S3.

Afterward, we incorporated names from resources such as WordNet (Brown 2005), Wikidata, DBpedia (Lehmann *et al.* 2015), ConceptNet (Speer *et al.* 2018), and 1085 existing

ontologies registered in BioPortal (Whetzel *et al.* 2011). LSF-related candidates were extracted from these resources and scored based on both semantic similarity to existing names and textual context using BERTopic (Grootendorst 2022). The details of this process are provided in Supplementary Section S4.

Figure 1 displays the details of different stages of LSF dictionary creation and expansion, along with methods used and the resources considered in each stage as LSF candidates. Apart from the initial creation of the dictionary, which was entirely performed manually, the subsequent two automatic expansion steps also included manual validation and filtering of candidates, before their addition to the dictionary.

2.4 Creation of an LSF ontology and referencing to external resources

In this study, we defined nine distinct categories, encompassing diverse LSF aspects, to establish a unified LSF classification across the various categories. For practical uses of the dictionary, such as concept indexing, it is important to know which names are synonyms for the same LSF and how the LSFs relate to each other. We thus created an LSF ontology (LSFO) in which each LSF is a concept with a unique identifier associated with relevant synonyms, and the concept can be traced back to the root of the LSFO through *is_a* relationships. To cover fine-grained concepts related to food groups, we imported the entire Food Groups branch of FoodOn, preserving its hierarchy. We also imported concepts from other ontologies; however, this was done in a selective manner involving manual curation and did not necessarily preserve the hierarchy of the source ontology. Details on the construction of LSFO and the process and tools used for conflict resolution are provided in Supplementary Section S5.

We utilized the BioPortal annotator to match the existing LSFs in our LSFO with names from over 1000 ontologies. To add cross-references (Xrefs) only to relevant ontologies, we relied on two metrics: *Coverage*, which determined the percentage of LSFs that exist in the target ontology, and *Overlap*, which assessed the percentage of names from the target ontology found within our LSFO. Based on these criteria, we manually narrowed down our selection to 50 ontologies to which we added Xrefs. This enhances interoperability and enables integration of existing ontologies, for example, by importing all child fine-grained concepts under the matched target name from domain-specific ontologies, furthering the development of a more comprehensive and interconnected LSFO.

The team of ontology curators and corpus annotators is a highly professional team with experience in the field. Specifically, six out of the eight members hold a PhD—and the remaining two an MSc—in bioinformatics or computer science. All authors work actively in the field of biomedical data science, half of them having biomedical NLP as their main research focus and having participated in the creation of several biomedical corpora (Kim *et al.* 2009, Pyysalo *et al.* 2012, Pafilis *et al.* 2013, Luoma *et al.* 2023, Mehryary *et al.* 2024, Nastou *et al.* 2024), and two having previously authored papers related to ontology design (Hoehndorf *et al.* 2011, Speer *et al.* 2018).

2.5 Dictionary-based NER

The JensenLab tagger (`tagger` hereafter) (https://github.com/larsjuhljensen/tagger) is a fast dictionary-based NER system that recognizes a wide variety of biomedical entities based on underlying dictionaries (Jensen 2016). To provide a dictionary-based NER system for LSF recognition, we integrated the `tagger` into our workflow, enhancing its functionality by supplying a dedicated LSF dictionary. The LSF dictionary was constructed using names from LSFO, alongside orthographic variant generation and a block list to enhance recall and precision, respectively. The `tagger` assigns a single unique identifier to synonyms and automatically generated name variations (e.g. plural and adjective forms) of the same LSF, enabling effective normalization for matched names. To improve precision, we use a block list to exclude problematic names that would cause many false positives during tagging. Specifically, we manually inspected all names that gave more than 2000 matches in 36.1 million PubMed abstracts (as of August 2023) and 4.5 million articles from the PMC open access subset (as of April 2022) to identify those that should be added to the block list.

2.6 Transformer-based NER

To explore the potential of using transformers for NER, we adapted an existing NER system (Luoma *et al.* 2023). Specifically, we built upon the RoBERTa-large-PM-M3-Voc model (RoBERTa-bio hereafter), which has demonstrated the best performance in several NER tasks (Lewis *et al.* 2020, Miranda-Escalada *et al.* 2023). We trained the model for multi-class classification of the nine categories of LSFs using LSF200 without OOC annotations. The 200 abstracts were divided into a 40-document holdout test set and a 160-document combined train and development set. We made sure this split was balanced across the nine categories of LSFO.

Hyperparameter selection was done through a grid search using a 5-fold cross-validation setup. This was done instead of a simple train–development split to avoid the risk of overfitting on a small development set. To determine the best

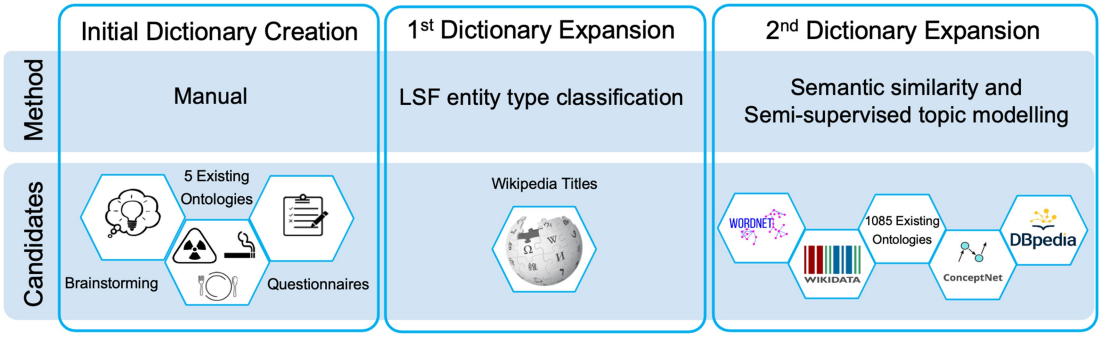

Figure 1. LSF dictionary creation and expansion stage.

hyperparameters for the model's final evaluation, we computed the micro F1-score based on the total true positives (TP), false positives (FP), and false negatives (FN) across all classes and folds. Finally, we trained a model on all available training and development data with the selected optimal hyperparameters and evaluated it on the holdout test set.

3 Results and discussion

3.1 The LSF ontology

Being able to identify and categorize concepts as diverse as LSFs was the main challenge of this work and it was achieved through the creation of a comprehensive and inclusive LSF ontology (LSFO). Table 1 presents an overview of LSFO, with the last column displaying the number of Xrefs per LSF category. One of the most important and well-studied categories of LSFs is Nutrition. To get more fine-grained concepts for this category, we incorporated food groups from the widely used FoodOn ontology. Table 1 presents both statistics with and without expansions from FoodOn.

3.2 The LSF200 corpus

LSF200 comprises 200 abstracts, with a total of 39 416 tokens based on BERT basic tokenization (https://github.com/spyysalo/bert-vocab-eval). It contains 1876 manually annotated mentions of LSFs, which are broken down into categories in Table 2. The majority of the LSF mentions fall into the categories *Nutrition, Socioeconomic factors, and Physical activities. In contrast, the categories Non-physical leisure time activities, Mental health practices, and Personal care products and cosmetic procedures* are less prevalent in LSF200. The distribution of names is consistent with the statistics of names in the different LSF categories of LSFO (Table 1), showcasing that despite selecting journals to represent all nine categories of LSFs, names from the largest categories tend to appear in most abstracts. Evaluating the quality of the manual annotations in terms of inter-annotator agreement gave an F1-score of 83%. This score emphasizes the challenging nature of annotating concepts as diverse as LSFs, making it difficult to formulate a comprehensive definition and achieve perfect annotation even at a human level.

We also introduced a special category of annotations called Out of context (OOC), which we use to highlight mentions that appear in a context that falls outside the scope of LSFs. For example, the word "tobacco" is normally an LSF but not when it appears in the context of "tobacco company."

3.3 Evaluation of dictionary-based NER

To assess the performance of our dictionary-based NER system, a crucial component involves evaluating the dictionary generated from the LSFO. It is important to note that this dictionary is not constructed using the text corpus, and, as such, it is appropriate to employ the entire LSF200 for evaluation of the dictionary and NER system.

We present performance results in two versions: One with OOC mentions included in the annotated corpus and one without. This evaluation assesses the impact of removing OOC mentions on the performance of dictionary-based NER. In the initial version, OOCs are treated as typical LSF mentions, resulting in an F1-score of 65.2% (precision: 96.0%, recall: 49.4%). In the second version, we removed OOC mentions entirely from annotations. The resulting performance was a slightly lower F1-score of 63.6% (precision: 85.0%, recall: 50.8%). When OOCs are not treated as LSFs, the number of false positives increases considerably, due to the presence of the OOC names in the dictionary, which results in a large drop in precision from 96.0% to 85.0%. Conversely, how OOC mentions are counted has very little impact on recall.

Figure 2 shows the performance for each of the nine categories, revealing that the recall varies much more than the precision and that OOC has a similar impact on precision across categories. For most categories, OOC mentions have only a minor impact on recall, with "non-physical time activities" being the only notable exception. For this category, many OOC mentions go undetected by the dictionary-based NER system, and the recall thus improves substantially when OOC mentions are excluded.

The wide range of recall values is explained by fundamental differences between the categories. For example, the dictionary by design does not have concepts for all possible exposures, for which reason the "Environmental exposures" category has very low recall. This implies that the low recall can likely be improved by using the existing Xrefs to integrate fine-grained concepts from, e.g. the Exposome Explorer.

In LSF200, this issue was not immediately evident, as it is selected from journals where all these entities are represented as LSFs. However, in a real-world scenario where the entire literature is tagged, this approach could result in detrimental effects on precision. Detailed performance results for various categories are available in Supplementary Section S6.

3.3.1 Manual error analysis

As anticipated, the dictionary-based NER exhibits impressive precision but lower recall due to its limitations in recognizing

Table 1. LSFC statistics.[a]

Category	Unique LSFs	Total names	Xrefs
Nutrition (expanded with FoodOn)	766 (4710)	987 (5126)	2093 (4875)
Socioeconomic factors	302	604	994
Environmental exposures	175	694	643
Substance use	104	283	596
Physical activities	84	171	283
Personal care products and cosmetic procedures	78	139	361
Non-physical leisure time activities	58	124	68
Sleep	45	104	129
Mental health practices	40	78	135
Total (expanded with FoodOn)	1652 (5597)	3184 (7324)	5302 (8097)

[a] In parentheses, the numbers including expansion with terms from FoodOn are shown for the *Nutrition* category and the total number of entries in LSFC.

Table 2. LSF200 corpus statistics.

Category	Total mentions	Unique names	Percent of names per category
Nutrition (expanded with FoodOn)	445	270	27.36
Socioeconomic factors	275	174	17.63
Environmental exposures	199	132	13.37
Substance use	187	75	7.60
Physical activities	222	88	8.92
Personal care products and cosmetic procedures	56	35	3.55
Non-physical leisure time activities	61	27	2.74
Sleep	110	57	5.78
Mental health practices	74	29	2.94
OOC	247	100	10.13
Total	1876	987	100

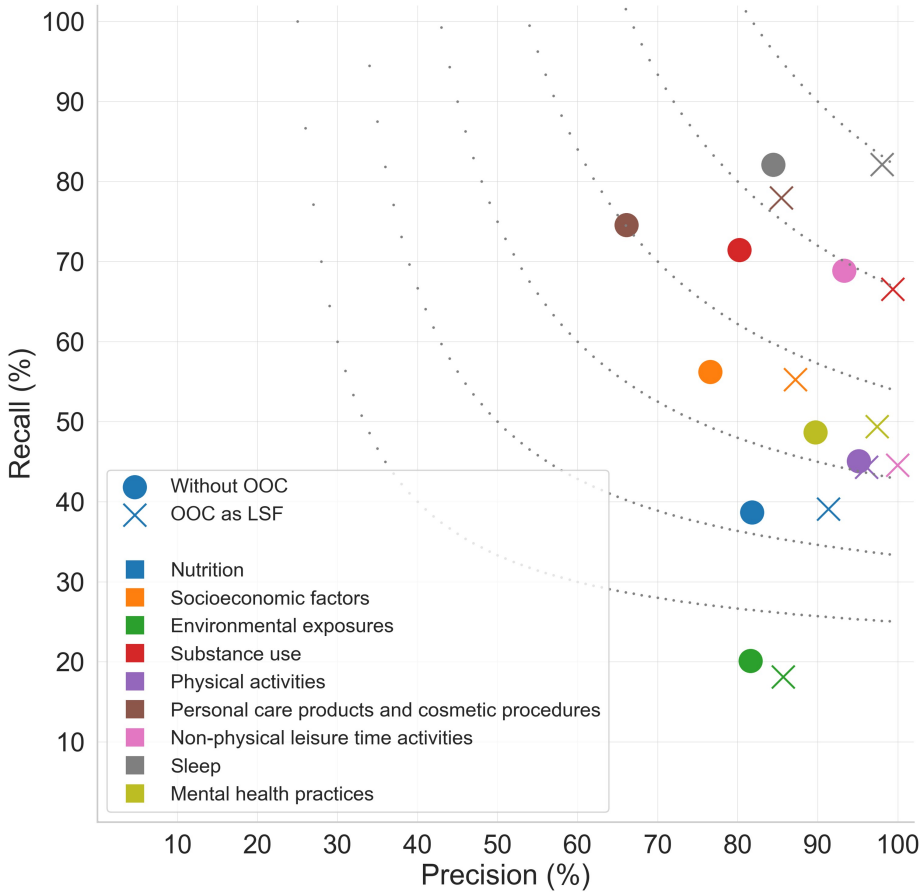

Figure 2. Performance of dictionary-based NER system in two OOC annotation variants per LSF category.

out-of-dictionary mentions. In Table 3, we have categorized the missing LSF mentions (false negatives) by the dictionary-based NER system.

The majority of missing LSF mentions are due to either abbreviations or synonyms related to existing LSF concepts in the dictionary. This is promising, as it suggests LSFO covers most of the relevant concepts, and that the recall of the NER system can thus be improved by enriching our dictionary with more synonyms for existing concepts. Other false negatives are due to ambiguous names that have been intentionally excluded; for example, "therapy" can denote both psychotherapy, which falls under mental health practices, or medical treatments, which are not considered LSFs. This approach prioritizes precision by avoiding matches with too many irrelevant hits, although it results in missing some relevant mentions in LSF200.

Table 3. Error analysis for dictionary-based NER (false negatives).

Error category	Total mentions	Unique names
Abbreviation and synonym	332	155
Ambiguous name	213	97
Fine-grained concept	211	128
Discontinuous entity	20	14
Annotation error	17	15
Total	793	409

The last major group of missing LSF mentions is "Fine-grained LSF concepts", which refers to mentions that cannot be classified as major missing LSF concepts because the parent LSFs of these mentions already exist in the dictionary. The "Environmental exposures" is, as already mentioned, a prominent example of this. The fact that missing concepts in LSFO account for only 211 errors in 1876 total LSF mentions, and that all of these are fine-grained concepts, suggests that the ontology has high quality in terms of its breadth.

The final few false negatives observed are either due to discontinuous entities, which cannot be identified by a dictionary-based method (e.g. "Vitamin C" in the expression "Vitamins E and C") or simply annotation errors made by the human curators and are thus not actual errors of the NER system.

As the dictionary-based system has high precision, it produces much fewer false positives than false negatives, the majority of which are due to OOC mentions as already described (Table 4). The remaining few false positives are either due to dictionary errors (names that should not have been in the dictionary), ambiguous names (which could be blocked at the price of more false negatives), or annotation errors in the corpus.

Table 4. Error analysis for dictionary-based NER (false positives).

Error category	Total mentions	Unique names
OOC	120	31
Ambiguous name	13	4
Annotation error	11	7
Dictionary error	3	2
Total	147	44

3.4 Comparison with transformer-based NER

To compare the performance of dictionary-based NER to state-of-the-art methods, we trained a transformer-based system on LSF200 without OOC mentions (see Methods for details). We evaluated the system on the 40-document test set of LSF200. The overall NER performance was 70.1% precision and 85.3% recall, corresponding to an F1-score of 77.0%. To allow direct comparison, we also reevaluated the dictionary-based NER system on the test set only, yielding 76.6% precision, 53.8% recall, and 63.2% F1-score. The results clearly highlight that the dictionary-based NER maximizes the precision by exclusively tagging predefined names from the dictionary and utilizing a block list to avoid tagging problematic names. In contrast, the transformer-based NER achieves better recall due to its ability to detect LSF mentions based on the context. Figure 3 shows the performance of dictionary-based NER and transformer-based NER for the test data within each LSF category.

The transformer-based NER demonstrates promising recall across most categories, with *Nutrition*, *Mental health practices*, and *Socioeconomic factors* showing particularly large improvements in recall with only small drops in precision. For *Personal care products and cosmetic procedures*,

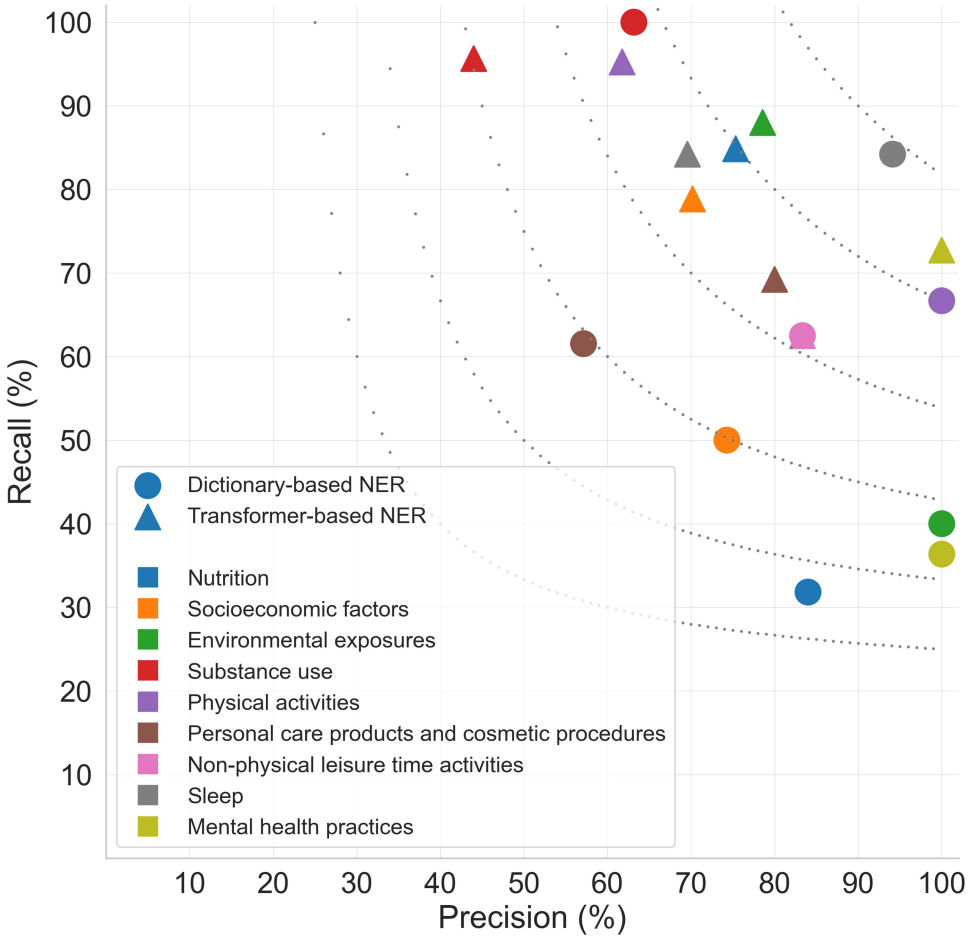

Figure 3. Performance of dictionary-based NER and transformer-based NER only for test data per LSF category.

Table 5. Error analysis for transformer-based NER.

	FP		FN	
Error category	**Total mentions**	**Unique names**	**Total mentions**	**Unique names**
Model error	40	23	31	25
OOC	25	6	0	0
Discontinuous entity	4	2	2	2
Annotation error	20	13	4	4
Total	89	44	37	31

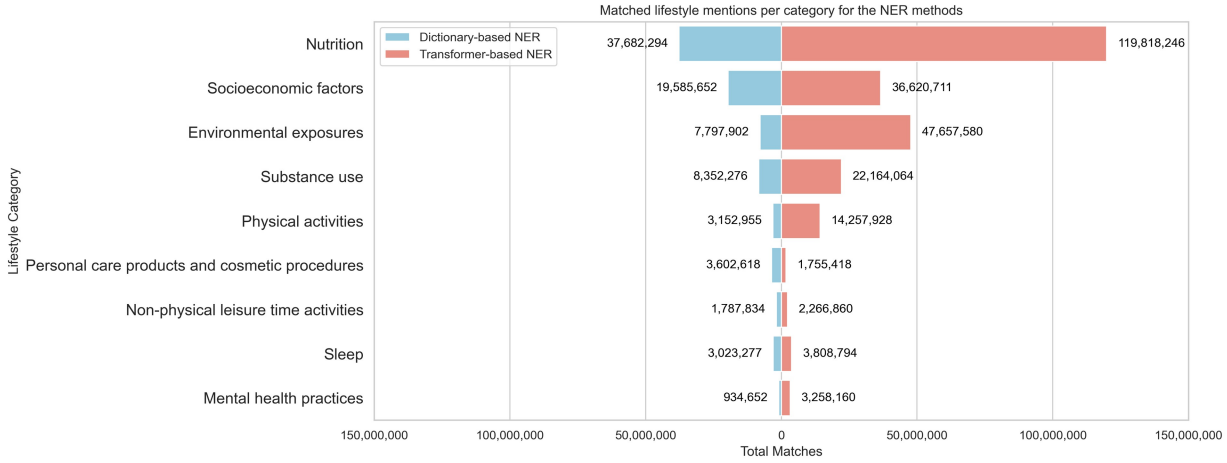

Figure 4. Matched lifestyle mentions from large-scale tagging of scientific literature using the NER methods.

transformer-based NER surprisingly improved mainly the precision. Finally, the transformer-based NER system performs worse than the dictionary-based one in terms of both precision and recall for categories *Substance use* and *Sleep*.

3.4.1 Manual error analysis

In Table 5, we categorize the errors produced by the Transformer-based NER, encompassing both names missed by the system (FNs) and names incorrectly detected as LSFs (FPs). The system also makes a few mistakes on discontinuous entities, although much fewer than the dictionary-based system.

The majority of the errors, however, were labeled as "Model error," encompassing the broad class of cases where the model, for no obvious specific reason, fails to provide accurate predictions. Lastly, we have again some errors, which upon inspection turn out to be mistakes in the manual annotation rather than the model being wrong.

3.5 Large-scale tagging of the scientific literature

Results for the tagging of 36.1 million PubMed abstracts (as of August 2023) and 4.5 million articles from the PMC open access subset (as of April 2022) using both the dictionary-based NER and the transformer-based NER have been made available via Zenodo (https://zenodo.org/records/10450308). There are in total 85 919 460 LSF matches for dictionary-based NER, with approximately half of them (44%) corresponding to *Nutrition* terms and a quarter (23%) to *Socioeconomic factors*. Tagging with the transformer-based system yielded 251 607 761 total matches. *Nutrition* terms once again constitute almost half of the matches of the system (48%), and *Environmental exposures* come second at 19%. Figure 4 shows the matches per LSF category for the two methods. A key difference between the results of the two systems is that the dictionary-based system inherently provides matches that are normalized to LSFO identifiers, whereas the transformer-based system does not. This makes the former a better starting point for many other text-mining tasks such as relation extraction.

4 Conclusions

In this paper, we introduce novel resources to address the challenge of recognizing LSFs within biomedical text. We present dictionary-based NER and transformer-based NER systems, both demonstrating promising performance in identifying LSFs. The LSF Classification stands out as a diverse and hierarchical classification of LSFs that we used as a backbone for the dictionary-based NER system, but which can serve as a resource for standardizing LSF information in general. Furthermore, the manually annotated LSF200 corpus proved to be sufficient for training a transformer-based NER and for evaluating both types of NER systems. The presented NER systems, LSFO, LSF200 corpus, and matched LSFs from large-scale runs of both NER systems on PubMed and PMC-OA articles, are made publicly available under open licenses to facilitate further research.

Supplementary data

Supplementary data are available at *Bioinformatics* online.

Conflict of interest

None declared.

Funding

This work was supported by the Novo Nordisk Foundation [NNF14CC0001, NFF17OC0027594]. K.N. has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie [101023676]. M.K. has received funding from Novo Nordisk Foundation [NNF20SA0035590].

References

Baron JA, Johnson CS-B, Schor MA *et al.* The DO-KB knowledgebase: a 20-year journey developing the disease open science ecosystem. *Nucleic Acids Res* 2024;**52**:D1305–14.

Brown K. Encyclopedia of Language & Linguistics. 2nd ed. ISBN: 9780080448541, Elsevier, 2005.

Buttigieg PL, Morrison N, Smith B *et al.*; ENVO Consortium. The environment ontology: Contextualising biological and biomedical entities. *J Biomed Semant* 2013;**4**:43.

Cook HV, Jensen LJ. A guide to dictionary-based text mining. In: Larson RS, Oprea TI (eds.), *Bioinformatics and Drug Discovery, Methods in Molecular Biology*. New York, NY: Springer, 2019, 73–89.

Dooley DM, Griffiths EJ, Gosal GS *et al.* FoodOn: a harmonized food ontology to increase global food traceability, quality control and data integration. *NPJ Sci Food* 2018;**2**:23.

Gabbert C, König IR, Lüth T *et al.* Lifestyle factors and clinical severity of Parkinson's disease. *Sci Rep* 2023;**13**:9537.

Gray ID, Kross AR, Renfrew ME *et al.* Precision medicine in lifestyle medicine: the way of the future? *Am J Lifestyle Med* 2020; **14**:169–86.

Grootendorst M. BERTopic: neural topic modeling with a class-based TF-IDF procedure. arXiv, arXiv:2203.05794, 2022, preprint: not peer reviewed.

Hoehndorf R, Ngonga Ngomo A-C, Pyysalo S *et al.* Ontology design patterns to disambiguate relations between genes and gene products in GENIA. *J Biomed Sem* 2011;**2**:S1.

Huang M-S, Lai P-T, Lin P-Y *et al.* Biomedical named entity recognition and linking datasets: Survey and our recent development. *Brief Bioinform* 2020;**21**:2219–38.

Jensen LJ. One tagger, many uses: illustrating the power of ontologies in dictionary-based named entity recognition. In: *Proceedings of the Joint International Conference on Biological Ontology and BioCreative*, Corvallis, OR: CEUR-WS.org, 2016.

Jeon J, Du M, Schoen RE *et al.*; Colorectal Transdisciplinary Study and Genetics and Epidemiology of Colorectal Cancer Consortium. Determining risk of colorectal cancer and starting age of screening based on lifestyle, environmental, and genetic factors. *Gastroenterology* 2018;**154**:2152–64.e19.

Kim J-D, Ohta T, Pyysalo S *et al.* Overview of BioNLP'09 shared task on event extraction. In: Tsujii J (ed.), *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*. Boulder, CO: Association for Computational Linguistics, 2009, 1–9.

Lehmann J, Isele R, Jakob M *et al.* DBpedia—a large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web* 2015; **6**:167–95.

Lewis P, Ott M, Du J *et al.* Pretrained Language Models for Biomedical and Clinical Tasks: Understanding and Extending the State-of-the-Art. In Rumshisky A, Roberts K, Bethard S, *et al.* (eds), *Proceedings of the 3rd Clinical Natural Language Processing Workshop*. Online: Association for Computational Linguistics, 2020, 146–157.

Luoma J, Nastou K, Ohta T *et al.* S1000: a better taxonomic name corpus for biomedical information extraction. *Bioinformatics* 2023; **39**:btad369.

Mehryary F, Nastou K, Ohta T *et al.* STRING-ing together protein complexes: corpus and methods for extracting physical protein interactions from the biomedical literature. *Bioinformatics* 2024;40:btae552.

Miranda-Escalada A, Mehryary F, Luoma J *et al.* Overview of DrugProt task at BioCreative VII: data and methods for large-scale text mining and knowledge graph generation of heterogenous chemical–protein relations. *Database* 2023;**2023**:baad080.

Nastou K, Koutrouli M, Pyysalo S *et al.* Improving dictionary-based named entity recognition with deep learning. *Bioinformatics* 2024;40:ii45–ii52.

Nastou K, Mehryary F, Ohta T *et al.* RegulaTome: a corpus of typed, directed, and signed relations between biomedical entities in the scientific literature. *Database* 2024;2024:baae095.

Neveu V, Nicolas G, Salek RM *et al.* Exposome-Explorer 2.0: an update incorporating candidate dietary biomarkers and dietary associations with cancer risk. *Nucleic Acids Res* 2020;**48**:D908–12.

Nyberg ST, Singh-Manoux A, Pentti J *et al.* Association of healthy lifestyle with years lived without major chronic diseases. *JAMA Intern Med* 2020;**180**:760–8.

Pafilis E, Frankild SP, Fanini L *et al.* The SPECIES and ORGANISMS resources for fast and accurate identification of taxonomic names in text. *PLoS One* 2013;**8**:e65390.

Perera N, Dehmer M, Emmert-Streib F *et al.* Named entity recognition and relation detection for biomedical information extraction. *Front Cell Dev Biol* 2020;**8**:673.

Pyysalo S, Ohta T, Rak R *et al.* Overview of the ID, EPI and REL tasks of BioNLP shared task 2011. *BMC Bioinform* 2012;**13**:S2.

Song B, Li F, Liu Y *et al.* Deep learning methods for biomedical named entity recognition: a survey and qualitative comparison. *Brief Bioinform* 2021;**22**:bbab282.

Speer R, Chin J, Havasi C. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. arXiv, arXiv:1612.03975, 2018, preprint: not peer reviewed.

Stenetorp P, Pyysalo S, Topić G *et al.* brat: a Web-based Tool for NLP-Assisted Text Annotation. In Segond F (ed), *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Avignon, France: Association for Computational Linguistics, 2012,102–107.

Subramanian M, Wojtusciszyn A, Favre L *et al.* Precision medicine in the era of artificial intelligence: Implications in chronic disease management. *JTransl Med* 2020;**18**:472.

Tobias DK, Merino J, Ahmad A *et al.* Second international consensus report on gaps and opportunities for the clinical translation of precision diabetes medicine. *Nat Med* 2023;**29**:2438–57.

Vaswani A, Shazeer N, Parmar N *et al.* Attention Is All You Need. arXiv, arXiv:1706.03762, 2017, preprint: not peer reviewed.

Whetzel PL, Noy NF, Shah NH *et al.* BioPortal: enhanced functionality via new web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. *Nucleic Acids Res* 2011;**39**:W541–545.

World Health Organization. Non communicable diseases. WHO 2023. https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases.

Younesi E, Ansari S, Guendel M *et al.* CSEO—the cigarette smoke exposure ontology. *J Biomed Semant* 2014;**5**:31.

Yurkovich JT, Evans SJ, Rappaport N *et al.* The transition from genomics to phenomics in personalized population health. *Nat Rev Genet* 2024;**25**:286–302.

Bioinformatics, 2024, 40, 1–8
https://doi.org/10.1093/bioinformatics/btae613
Original Paper